Single Channel Informed Signal Separation using Artificial-Stereophonic Mixtures and Exemplar-Guided Matrix Factor Deconvolution

Ahmed Al-Tmeme[1], W.L. Woo[2,*], S.S. Dlay[2], Bin Gao[3]

[1]Information and Communication Engineering Department, University of Baghdad, Iraq
[2]School of Electrical and Electronic Engineering, Newcastle University, England, United Kingdom
[3]School of Automation Engineering, University of Electronic Science and Technology of China, China
[*]Corresponding e-mail: lok.woo@ncl.ac.uk

***Abstract* — In this paper, a method is proposed to tackle the problem of single channel audio separation. The proposed method leverages on the exemplar source is used to emulate the targeted speech signal. A multi-component nonnegative matrix factor 2-D deconvolution (NMF2D) is proposed to model the temporal and spectral changes, and the number of spectral basis of the audio signals. The paper proposes an artificial auxiliary channel to imitate a pair of stereo mixture signals which is termed as "artificial-stereophonic mixtures". The artificial-stereophonic mixtures and the exemplar source are jointly used to guide the factorization process of the NMF2D. The factorization is adapted under a hybrid framework that combines the Generalized Expectation-Maximization algorithm with multiplicative update adaptation. The proposed algorithm leads to fast and stable convergence, and ensures the non-negativity constraints of the solution are satisfied. Adaptive sparsity has also been introduced on each sparse parameter in the multi-component NMF2D model when the exemplar deviates from the target signal. Experimental results have shown the competence of the proposed algorithms in comparison with other algorithms.**

***Keywords* — adaptive signal processing, signal processing, learning systems, intelligent control, system identification**

I. INTRODUCTION

Blind source separation (BSS) [1-5] is an ill-posed problem that cannot be totally solved without some prior information, i.e., a certain number of assumptions have to be imposed to render the problem solvable [2] such as mutual statistical independence among the sources, the number of sources, how the sources are mixed, the location of the sources with respect to the microphones, and the channel type. Several recent solutions have been developed to mitigate some of these constraints. In the work [6], a method is proposed to decorrelate multiple non-stationary stochastic sources using a multivariable crosstalk-resistant adaptive noise canceller. In

related method [7], the problem of speech quality enhancement is tackled using adaptive and non-adaptive filtering algorithms. A two-microphone Gauss-Seidel pseudo affine projection algorithm combined with forward blind source separation is proposed. A higher efficiency in speech enhancement in noisy environment has been attained. The paper [8] proposes rational polynomial functions to replace the original score functions used in standard ICA. The rational polynomials are derived by the Pade approximant from Taylor series expansion of the original nonlinearities which can be quickly evaluated to enable large-scale multidimensional sets of data characterized by super-Gaussian distribution to be separated within a short period of time.

In recent years, there has been a shift of focus from blind to informed source separation framework where the aim is to achieve higher performance that the conventional BSS approaches cannot reach. In this framework, researchers seek an aid from an external source in addition to the mixture signal as side information to enhance the separation performance. Informed source separation can be classified as follows:

(a) Score Informed Source Separation: In this method the parameters of the separation algorithm are initialized by depending on the side information that are available from the Musical Instrument Digital Interface (MIDI) files (sometimes they are called musical scores), such as the onset time, pitch, and duration of the musical notes [9-10]. An overview of the score informed source separation can be found in the paper [11]. Furthermore, similar to this idea the user can manually set or rest the code matrix in the NMF models [12-13].

(b) Exemplar-Based Source Separation: Here the informed source separation targeted a specific source in the mixture by providing another source that is similar to the one to be separated. Such as the user mimic the targeted source by singing [14], by humming [15], or by dubs the dialog in films [16]. Furthermore, using an additional audio references as a side information such as using the multitrack cover version of the same song [17-20] or using several international versions of the same movie [21]. Additionally the text can be used as side information to mimic the targeted speech signal [22]. The use of the exemplar gives a controllability over the separation mechanism by mean of which source is to be separated in the mixture [23-25].

In both approaches, there is a need for a synthesizer to convert them to music. In the score informed source separation an MIDI synthesizer or a user is usually used to convert the scores to music in order to use them as side information with the audio mixture. Similarly, the exemplar-based informed source separation (especially text based one) uses a speech synthesizer or a user to convert the texts to music.

(c) Coding Based Informed Source Separation: It is two stages scenario that contains the encoding stage and the decoding stage. At the encoding stage all the sources are available in addition to the mixture in order to generate a side information that can be transmitted with the mixture or be embedded in the mixture [26], and will be used in the decoder stage to separate the sources [27-30]. Ozerov *et al.* [30] showed that the coding based informed source separation can outperforms the oracle estimation, if the required bitrate provided. It is

bitrate vs quality of separation in this type of informed source separation, as it takes advantage from both source coding and source separation.

Among these types of informed source separation, the exemplar based informed source separation has been pursued in this paper as the MIDI files are not always available in the case of the score informed source separation. Also, the coding based informed source separation did not progress far as it investigates the quality of separation achieved in terms of the available bitrate, and therefore it is far from the scope covered of this paper; however it is very prompting future work if it can be proven that the nonnegative matrix factor 2-D deconvolution (NMF2D) [31] can achieve better performance and lower bitrate than the conventional methods.

In this paper, we address the problem of informed source separation when only a single receiver is available. This gives rise to the single-channel source separation (SCSS) which is an extreme case of under-determined BSS problem where only a single channel recording is available to estimate more than one source signals. The proposed solution leverages on exemplar signal from text associated with the mixture which is generated by using a speech synthesizer or human speaker. The approach is essentially belonging to the category of text informed source separation [22, 32-33]. In the current paper, two algorithms will be proposed, namely, the full exemplar-based algorithm and the semi-exemplar based algorithm[*]. In the exemplar-based algorithm, we will utilize the exemplar to "guide" the factorization process of the NMF2D. The cost function is augmented with a weighted term containing the exemplar which is co-factorized along with the mixture signal. We term this as the Nonnegative Matrix Partial Co-factorization 2-D Deconvolution (NMPCF2D). The proposed NMPCF2D has the ability to describe pitch and temporal changes of the signal as well as the variations in the spectral basis. In the case of semi-exemplar based algorithm, the exemplar signal is used only to initialize the tensors of the NMF2D which alone will be used to carry out the separation. The difference between the semi-exemplar based algorithm and the exemplar based algorithm is that the former algorithm will guide the separation for the first iteration only (i.e., to give the correct start) by initializing its tensors through the exemplar signal, while in the latter algorithm the exemplar signal is used to initialize as well as to guide the separation process for every iteration via the NMPCF2D until it converges to the desired solution. For faster convergence, both algorithms are adapted under a framework that hybridizes the Generalized EM algorithm with the Multiplicative Update (which is termed as GEM-MU [12]). Furthermore, as the speech source changes rapidly over time, conventional methods that assign a fixed uniform sparsity will inadvertently lead to either too many ineffective temporal codes (i.e. under-sparseness), or too many temporal codes set to zero (i.e. over-sparseness). Thus the proposed solution addresses this issue by internalizing a mechanism wherein the sparsity level of each individual temporal code is adaptively tuned as part of the parameter estimation stage. Finally, to relieve the ill-posed condition imposed by the single recording of the mixture signal, the paper proposes an "auxiliary channel"

[*] For brevity, we term the full-exemplar based algorithm simply as exemplar based algorithm.

which generates another mixture signal. When this auxiliary mixture signal concatenates with the original mixture, they emulate a pair of stereophonic mixture signals. Fig. 1 shows a high level representation of the proposed exemplar-based algorithm. The semi-exemplar algorithm is similar to Fig. 1 except that the exemplar signal does not directly feed into the GEM-MU based NMPCF2D.

The contributions in this paper can be summarized as follows: Firstly, two informed source separation algorithms are proposed. Secondly, the proposal of NMPCF2D to guide the source separation process. Thirdly, the development of artificial stereophonic channel and its integration with the hybrid GEM-MU algorithm to render stable and fast convergence during parameter estimation while preserving the non-negativity constraints of the sources. The paper is organized as follows: Section II is dedicated to the problem formulation where the mixture model with artificial stereophonic channel and the exemplar model will be formulated. The proposed full-exemplar and semi-exemplar based algorithms will be derived in Section III. The description of the targeted speech signal using the exemplar signal will be carried out in Section IV. Experimental results and discussions of these results will be shown in Section V. Finally, Section VI draws the conclusions.

II. Problem Formulation

A. Auxiliary Channel

Considering the underdetermined single channel mixture, namely:

$$\tilde{x}_1(t) = \tilde{g}(t) + \tilde{b}(t) + \tilde{n}(t) \tag{1a}$$

where $x_1(t)$ is the sampled mixture signal, $\tilde{g}(t)$ the sampled target signal (e.g. speech), and $\tilde{b}(t)$ is the sampled background signal (which we will take as either a music or effects (fx)), $\tilde{n}(t)$ is some additive noise, for $(t = 1, \dots, T)$. In this paper, $\tilde{g}(t)$ and $\tilde{b}(t)$ are assumed to be modelled by the autoregressive (AR) process:

$$\tilde{g}(t) = -\sum_{\tau=1}^{D_{\tilde{g}}} a_{\tilde{g}}(\tau)\tilde{g}(t-\tau) + e_{\tilde{g}}(t) \tag{1b}$$

where $a_{\tilde{g}}(\tau)$ denotes the τ^{th} order AR coefficient of target signal, $D_{\tilde{g}}$ is the maximum AR order, and $e_{\tilde{g}}(t)$ is an independent identically distributed (i.i.d.) random signal with zero mean and variance $\sigma^2_{e_{\tilde{g}}}$. The same definition will be applied to the background signal i.e. $\tilde{b}(t) = -\sum_{\tau=1}^{D_{\tilde{b}}} a_{\tilde{b}}(\tau)\tilde{b}(t-\tau) + e_{\tilde{b}}(t)$. This model is particularly interesting in signal separation. Firstly, many real-life signals satisfy this process and secondly, it enables us to formulate an auxiliary channel by weighting and time-shifting $x_1(t)$ as

$$\tilde{x}_2(t) = \frac{\tilde{x}_1(t) + \gamma\tilde{x}_1(t-\delta)}{1+|\gamma|} \tag{2}$$

In Eqn. (2), $\gamma \in \Re$ is the weight parameter, and δ is the time-delay between $\tilde{x}_2$ and $\tilde{x}_1$. The original mixture in Eqn. (1a) together with the auxiliary channel output in Eqn. (2) form what we term in this paper as

"artificial-stereophonic mixtures". It has an artificial resemblance of a stereo signal except that it is given at one spatial location which results in the same time-delay but different attenuation of the source signals. To show this, we can express Eqn. (2) in terms of the source signals, AR coefficients and time-delay as

$$\tilde{x}_2(t) = \frac{\tilde{g}(t) + \tilde{b}(t) + \tilde{n}(t) + \gamma\left[\tilde{g}(t-\delta) + \tilde{b}(t-\delta) + \tilde{n}(t-\delta)\right]}{1+|\gamma|}$$

$$= \frac{\left(-a_{\tilde{g}}(\delta) + \gamma\right)}{1+|\gamma|}\tilde{g}(t-\delta) + \frac{(-a_{\tilde{b}}(\delta) + \gamma)}{1+|\gamma|}\tilde{b}(t-\delta) + \frac{\tilde{n}(t) \; + \; \gamma\tilde{n}(t-\delta)}{1+|\gamma|}$$

$$+ \frac{e_{\tilde{g}}(t) - \sum_{\substack{\tau=1 \\ \tau\neq\delta}}^{D_{\tilde{g}}} a_{\tilde{g}}(\tau)\tilde{g}(t-\tau)}{1+|\gamma|} + \frac{e_{\tilde{b}}(t) - \sum_{\substack{\tau=1 \\ \tau\neq\delta}}^{D_{\tilde{b}}} a_{\tilde{b}}(\tau)\tilde{b}(t-\tau)}{1+|\gamma|}. \quad (3)$$

Defining the followings:

$$a_{\tilde{g}}(\delta) = \frac{-a_{\tilde{g}}(\delta)+\gamma}{1+|\gamma|}, \; a_{\tilde{b}}(\delta) = \frac{-a_{\tilde{b}}(\delta)+\gamma}{1+|\gamma|}$$

$$r_{\tilde{g}}(t) = \frac{e_{\tilde{g}}(t) - \sum_{\substack{\tau=1 \\ \tau\neq\delta}}^{D_{\tilde{g}}} a_{\tilde{g}}(\tau)\tilde{g}(t-\tau)}{1+|\gamma|}$$

$$r_{\tilde{b}}(t) = \frac{e_{\tilde{b}}(t) - \sum_{\substack{\tau=1 \\ \tau\neq\delta}}^{D_{\tilde{b}}} a_{\tilde{b}}(\tau)\tilde{b}(t-\tau)}{1+|\gamma|}$$

$$\nu(t) = r_{\tilde{g}}(t) + r_{\tilde{b}}(t) + \frac{\tilde{n}(t) \; + \; \gamma\tilde{n}(t-\delta)}{1+|\gamma|}. \quad (4)$$

In Eqn. (4), $r_j(t)$ represents the residue of the speech signal and background, and $\nu(t)$ denotes the accumulated noise obtained by weighting and time-shifting of the additive noise $\tilde{n}(t)$ plus the residues. Using Eqns. (1)-(4), the mixture model can now be formulated in terms of the sources and the noise as

$$\tilde{x}_1(t) = \tilde{g}(t) + \tilde{b}(t) + \tilde{n}(t)$$
$$\tilde{x}_2(t) = a_{\tilde{g}}(\delta)\tilde{g}(t-\delta) + a_{\tilde{b}}(\delta)\tilde{b}(t-\delta) + \nu(t). \quad (5)$$

Eqn. (5) shows that the signals $\tilde{x}_1(t)$ and $\tilde{x}_2(t)$ *resemble* a pair of stereo mixture signals that has been mixed by $\tilde{g}(t)$ and $\tilde{b}(t)$ and their constituents delayed version. The terms $\tilde{n}(t)$ and $\nu(t)$ represent the noise that perturbed $\tilde{x}_1(t)$ and $\tilde{x}_2(t)$, respectively.

The time-frequency representation of the noisy mixing model is obtained using the Short-Time Fourier Transform (STFT) of $\tilde{x}_j$, $j = 1,2$ as

$$x_{1,f,n} = g_{f,n} + b_{f,n} + n_{f,n}$$
$$x_{2,f,n} = a_{\tilde{g}}(\delta)e^{-i\omega\delta}g_{f,n-\delta} + a_{\tilde{b}}(\delta)e^{-i\omega\delta}b_{f,n-\delta} + \nu_{f,n} \quad (6)$$

where $x_{i,f,n}$, $g_{f,n}$, $b_{f,n}$, $n_{f,n}$ and $\nu_{f,n}$ are the STFT of $\tilde{x}_i(t)$, $\tilde{g}(t)$, $\tilde{b}(t)$, $\tilde{n}(t)$ and $\nu(t)$, respectively, and $f = 1, \ldots, F$ is the frequency bin index. By invoking stationarity of the source signals i.e., $STFT[\tilde{g}(t-\delta)]$

$= e^{-i\omega\delta} g_{f,n-\delta} \approx e^{-i\omega\delta} g_{f,n}$, the above can be expressed as

$$\boldsymbol{x}_{f,n} \cong \boldsymbol{A}_f \boldsymbol{s}_{f,n} + \mathbf{n}_f \tag{7}$$

where $\boldsymbol{x}_{f,n} = \begin{bmatrix} x_{1,f,n} \\ x_{2,f,n} \end{bmatrix} \in \mathbb{C}^{2\times 1}$, $\boldsymbol{A}_f = \begin{bmatrix} 1 & 1 \\ a_{\tilde{g},f}(\delta) & a_{\tilde{b},f}(\delta) \end{bmatrix} \in \mathbb{C}^{2\times 2}$, $a_{j,f}(\delta) = a_j(\delta)e^{-i\omega\delta}$, $\boldsymbol{s}_{f,n} = \begin{bmatrix} g_{f,n} \\ b_{f,n} \end{bmatrix} \in \mathbb{C}^{2\times 1}$, and $\mathbf{n}_f = \begin{bmatrix} n_{f,n} \\ v_{f,n} \end{bmatrix} \in \mathbb{C}^{2\times 1}$. The separability of the proposed pseudo-stereo mixture has been undertaken and presented in the Appendix. It is shown that when the AR coefficients of the sources at selected delay are different (i.e. $a_{\tilde{g}}(\delta) \neq a_{\tilde{b}}(\delta)$) or the residues are different (i.e. $r_{\tilde{g}}(t) \neq r_{\tilde{b}}(t)$), then the pseudo-stereo mixture in (7) is separable. It should be noted that through delaying the original mixture $\tilde{x}_1(t)$ by δ lag and weightily recombining it with the original mixture in Eqn. (2), the resulting action resembles a filtering process by a finite impulse response filter. As a result, the waveform of the auxiliary channel output differs from the original mixture. Also since the sources are audio signals, the local stationarity assumption enables the AR coefficients at the corresponding time delay to manifest as if these were the "mixing coefficients". Hence this enables a representation of two-input two-output mixing system as shown in Eqn. (7). These mixing coefficients culminated to a square matrix which has the attribute of a full-rank matrix provided that $a_{1,f}(\delta) \neq a_{2,f}(\delta)$. Thus the single-channel source separation is transformed into an exact-determined system in the TF domain given the above conditions hold. To further model the sources $\boldsymbol{s}_{f,n}^T = [g_{f,n} \quad b_{f,n}]$, we propose to use a multi-component NMF2D. The choice of NMF2D is motivated by the need to specify the spectral and temporal changes of the target speech signal through its convolutive parameters and the number of frequency basis. If NMF is used, it will be able to describe the number of frequency basis only. Therefore, the NMF2D with multiple components will be considered for the spectral variance model instead of the NMF spectral model. Thus, each source can be expressed by K complex-valued latent components, i.e., $g_{f,n} = \sum_{k=1}^{K_g} c_{k,f,n}^g$ and $b_{f,n} = \sum_{k=1}^{K_b} c_{k,f,n}^b$ and can be modeled as realization of proper complex zero-mean variables:

$$\begin{aligned} c_{k,f,n}^g &\sim \mathcal{N}_c\left(0, \sigma_{k,f,n}^{g^2}\right) = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{k,f-\phi,\tau}^g h_{k,\phi,n-\tau}^g\right) \\ c_{k,f,n}^b &\sim \mathcal{N}_c\left(0, \sigma_{k,f,n}^{b^2}\right) = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{k,f-\phi,\tau}^b h_{k,\phi,n-\tau}^b\right) \end{aligned} \tag{8}$$

where $\mathcal{N}_c(\mu, \Sigma)$ is proper complex Gaussian distribution [34], $w_{k,f,\tau}^g$ and $w_{k,f,\tau}^b$ represent the spectral basis of the speech and background sources, respectively; $h_{k,\phi,n}^g$ and $h_{k,\phi,n}^b$ represent the temporal code for each spectral basis element of the speech and background sources, respectively, for $f = 1, \ldots, F;\ n = 1, \ldots, N;\ k = 1, \ldots K$.

B. Generalized Expectation-Maximization (GEM) Algorithm

The maximum *a posteriori* (MAP) probability is chosen as the optimization criterion. The noise $n_{i,fn}$ is assumed to be stationary and spatially uncorrelated, i.e.

$$n_{i,fn} \sim \mathcal{N}_c\left(0, (\sigma_{i,f}^n)^2\right) \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{n},f} = diag\left[(\sigma_{i,f}^n)^2\right]. \tag{9}$$

Let $\boldsymbol{X} = \{\boldsymbol{x}_{f,n}\}_{f,n}$ and $\boldsymbol{C} = \left\{\left\{c_{k,f,n}^g\right\}, \{c_{k,f,n}^b\}\right\}_{k,f,n}$ be the observations and latent variables, and $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\mathbf{n}}\}$ as the parameters of the model where $\boldsymbol{A} = \{\boldsymbol{A}_f\}_f$, $\boldsymbol{W} = \{\boldsymbol{W}^g, \boldsymbol{W}^b\}$, $\boldsymbol{H} = \{\boldsymbol{H}^g, \boldsymbol{H}^b\}$, $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}^g, \boldsymbol{\Lambda}^b\}$, $\boldsymbol{W}^g = \left\{w_{k,f,\tau}^g\right\}_{k,f,\tau}$, $\boldsymbol{W}^b = \{w_{k,f,\tau}^b\}_{k,f,\tau}$, $\boldsymbol{H}^g = \left\{h_{k,\phi,n}^g\right\}_{k,\phi,n}$, $\boldsymbol{H}^b = \{h_{k,\phi,n}^b\}_{k,\phi,n}$, $\boldsymbol{\Lambda}^g = \left\{\lambda_{k,\phi,n}^g\right\}_{k,\phi,n}$, $\boldsymbol{\Lambda}^b = \{\lambda_{k,\phi,n}^b\}_{k,\phi.n}$. The tensor $\boldsymbol{\Lambda}$ contains the sparsity terms for $\boldsymbol{H}$. The estimation of model parameters and latent variables will alternate between the expectation-maximization (EM) steps through the posterior probability:

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\theta} \log p(\boldsymbol{\theta}|\boldsymbol{X})$$

where

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}) \geq \int Q(\boldsymbol{C}) \log\left[\frac{p(\boldsymbol{C}, \boldsymbol{\theta}|\boldsymbol{X})}{Q(\boldsymbol{C})}\right] d\boldsymbol{C} \tag{10}$$

for any distribution $Q(\boldsymbol{C})$. Defining $F(Q(\boldsymbol{C}), \boldsymbol{\theta}) = \int Q(\boldsymbol{C}) \log\left[\frac{p(\boldsymbol{C},\boldsymbol{\theta}|\boldsymbol{X})}{Q(\boldsymbol{C})}\right] d\boldsymbol{C}$, then the E-step consists of determining $Q(\boldsymbol{C})$ that maximizes $F(Q(\boldsymbol{C}), \boldsymbol{\theta})$ where the optimal $Q(\boldsymbol{C})$ is given by $Q^*(\boldsymbol{C}) = P(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{\theta}')$ for the current model $\boldsymbol{\theta}'$. The M-step consists of maximizing $F(Q^*(\boldsymbol{C}), \boldsymbol{\theta})$ with respect to the model $\boldsymbol{\theta}$ when $Q(\boldsymbol{C})$ is fixed at $Q^*(\boldsymbol{C})$ i.e. $\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \int Q^*(\boldsymbol{C}) \log p(\boldsymbol{C}, \boldsymbol{\theta}|\boldsymbol{X})\, d\boldsymbol{C}$. The posterior probability is given by

$$p(\boldsymbol{C}, \boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}, \boldsymbol{C}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \propto p(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{\theta}) p(\boldsymbol{C}|\boldsymbol{\theta}) P(\boldsymbol{\theta}). \tag{11}$$

III. Proposed Exemplar and Semi-Exemplar Algorithms

The GEM-MU will be used as the platform for deriving the proposed algorithms. The GEM-MU combines the generalized EM algorithm and the multiplicative update (MU) algorithm. The source power spectrogram posterior estimates ($\hat{p}_{j,fn}$) (see Eqn. (13)), the mixing parameters, and the noise covariance will be estimated in the E-step of the EM algorithm, while the parameters W and H will be estimated in the M-step of the EM algorithm by using the MU algorithm with adaptive sparsity NMF2D. First of all, the common part between the two proposed algorithms will be derived, and once this is achieved we will derive each one separately within their context.

A. E-Step: Conditional expectations of natural statistics

In the E-step, the complete data $\{\boldsymbol{X}, \boldsymbol{C}\}$ and its pdfs $p(\boldsymbol{X}, \boldsymbol{C}|\boldsymbol{\theta})$ form an exponential family. Using Eqns. (7)-(9), it can be shown that the complete data log-likelihood is given by

$$
\begin{aligned}
-\log p(\boldsymbol{C}, \boldsymbol{\theta}|\boldsymbol{X}) &= -\log p(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{\theta}) - \log p(\boldsymbol{C}|\boldsymbol{\theta}) - \log P(\boldsymbol{\theta}) \\
&\stackrel{c}{=} \sum_{f,n}\Bigg[\log\left|\boldsymbol{\Sigma}_{\mathbf{n},f}\right| + \sum_{k=1}^{K_g}\log\left(\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w^{g}_{k,f-\phi,\tau} h^{g}_{k,\phi,n-\tau}\right) + \sum_{k=1}^{K_g}\frac{\left|c^{g}_{k,f,n}\right|^2}{\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w^{g}_{k,f-\phi,\tau} h^{g}_{k,\phi,n-\tau}} \\
&\quad + \sum_{k=1}^{K_b}\log\left(\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w^{b}_{k,f-\phi,\tau} h^{b}_{k,\phi,n-\tau}\right) + \sum_{k=1}^{K_b}\frac{\left|c^{b}_{k,f,n}\right|^2}{\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w^{b}_{k,f-\phi,\tau} h^{b}_{k,\phi,n-\tau}}\Bigg] \\
&\quad + N\sum_{f} tr\left[\boldsymbol{\Sigma}^{-1}_{\mathbf{n},f}\boldsymbol{R}_{\boldsymbol{xx},f} - \boldsymbol{\Sigma}^{-1}_{\mathbf{n},f}\boldsymbol{A}_f\boldsymbol{R}^{H}_{\boldsymbol{xs},f} - \boldsymbol{\Sigma}^{-1}_{\mathbf{n},f}\boldsymbol{R}_{\boldsymbol{xs},f}\boldsymbol{A}^{H}_f + \boldsymbol{\Sigma}^{-1}_{\mathbf{n},f}\boldsymbol{A}_f\boldsymbol{R}_{\boldsymbol{ss},f}\boldsymbol{A}^{H}_f\right] \\
&\quad - \log p\left(\boldsymbol{A}_f\right) - \log p\left(\boldsymbol{\Sigma}_{\mathbf{n},f}\right) - \log p(\boldsymbol{W}) - \log p(\boldsymbol{H}|\boldsymbol{\Lambda})
\end{aligned} \tag{12}
$$

where $\boldsymbol{R}_{\boldsymbol{xx},f} = \frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{x}^{H}_{fn}$, $\boldsymbol{R}_{\boldsymbol{ss},f} = \frac{1}{N}\sum_n \boldsymbol{s}_{fn}\boldsymbol{s}^{H}_{fn}$ and $\boldsymbol{R}_{\boldsymbol{xs},f} = \frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{s}^{H}_{fn}$. The above complete data likelihood can be represented in the standard form of exponential family. In other words, we have $\log p(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{\theta}) = \langle \eta(\boldsymbol{\theta}), \mathrm{T}(\boldsymbol{X}, \boldsymbol{C})\rangle + \vartheta(\boldsymbol{\theta})$ where $\mathrm{T}(\boldsymbol{X}, \boldsymbol{C})$ is a vector of all scalar elements of $\mathrm{t}(\boldsymbol{X}, \boldsymbol{C}) = \left\{\boldsymbol{R}_{\boldsymbol{xx},f}, \boldsymbol{R}_{\boldsymbol{ss},f}, \boldsymbol{R}_{\boldsymbol{xs},f}, u^{g}_{k,f,n}, u^{b}_{k,f,n}\right\}$ where $u^{j}_{k,f,n} = \left|c^{j}_{k,f,n}\right|^2$ for $j = \{g, b\}$, and $\eta(\boldsymbol{\theta})$ and $\vartheta(\boldsymbol{\theta})$ are some vector and scalar functions of parameters. The natural sufficient statistics of this family is given by $\mathrm{t}(\boldsymbol{X}, \boldsymbol{C})$. In the E-step, the conditional expectation of the natural statistics are evaluated according to $\hat{\mathrm{t}}(\boldsymbol{X}, \boldsymbol{\theta}') = \int \mathrm{t}(\boldsymbol{X}, \boldsymbol{C})\, p(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{\theta}')d\boldsymbol{C}$ and this gives the conditional expectations $\widehat{\boldsymbol{R}}_{\boldsymbol{xx},f} = \boldsymbol{R}_{\boldsymbol{xx},f} = \frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{x}^{H}_{fn}$, $\widehat{\boldsymbol{R}}_{\boldsymbol{xs},f} = \frac{1}{N}\sum_n \boldsymbol{x}_{fn}\hat{\boldsymbol{s}}^{H}_{fn}$, $\widehat{\boldsymbol{R}}_{\boldsymbol{ss},f} = \frac{1}{N}\sum_n \hat{\boldsymbol{s}}_{fn}\hat{\boldsymbol{s}}^{H}_{fn} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{s},fn}$ and $\hat{u}^{j}_{k,f,n} = \left[\hat{\boldsymbol{c}}_{fn}\hat{\boldsymbol{c}}^{H}_{fn} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{c},fn}\right]^{j}_{k,k}$ where $\hat{\boldsymbol{s}}_{fn} = \langle \boldsymbol{s}_{fn}|\boldsymbol{x}_{fn}, \boldsymbol{\theta}'\rangle$ is the *a posteriori* estimate of the source using the model θ' estimated from previous EM step. The source power spectrogram posterior estimates are given by

$$\hat{p}_{j,fn} = \hat{R}_{\boldsymbol{ss},fn}(j, j) \tag{13}$$

where

$$\hat{\boldsymbol{s}}_{fn} = \boldsymbol{\Sigma}_{\boldsymbol{s},fn}\boldsymbol{A}^{H}_f\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x},fn}\boldsymbol{x}_{fn} \tag{14}$$

$$\hat{\boldsymbol{c}}_{fn} = \boldsymbol{\Sigma}_{\boldsymbol{c},fn}\left[\boldsymbol{A}_f \otimes \mathbf{1}_K\right]^{H}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x},fn}\boldsymbol{x}_{fn} \tag{15}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{x},fn} = \boldsymbol{A}_f\boldsymbol{\Sigma}_{\boldsymbol{s},fn}\boldsymbol{A}^{H}_f + \boldsymbol{\Sigma}_{\mathbf{n},f} \tag{16}$$

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{s},fn} = \left(\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{s},fn}\boldsymbol{A}^{H}_f\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x},fn}\boldsymbol{A}_f\right)\boldsymbol{\Sigma}_{\boldsymbol{s},fn} \tag{17}$$

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{c},fn} = \left(\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{c},fn}\left[\boldsymbol{A}_f \otimes \mathbf{1}_K\right]^{H}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x},fn}\left[\boldsymbol{A}_f \otimes \mathbf{1}_K\right]\right)\boldsymbol{\Sigma}_{\boldsymbol{c},fn} \tag{18}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{s},fn} = \begin{bmatrix} \sum_{k,\tau,\phi} w^{g}_{k,f-\phi,\tau} h^{g}_{k,\phi,n-\tau} & 0 \\ 0 & \sum_{k,\tau,\phi} w^{b}_{k,f-\phi,\tau} h^{b}_{k,\phi,n-\tau} \end{bmatrix} \tag{19}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{c},fn} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{c}^{g},fn} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{c}^{b},fn} \end{bmatrix} \tag{20a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{c}^{j},fn} = diag\left(\left[\sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w^{j}_{k,f-\phi,\tau} h^{j}_{k,\phi,n-\tau} \right]_{k,k} \right). \tag{20b}$$

In above, '$\otimes$' is the Kronecker product and $\mathbf{1}_K$ is a row vector with K unit element where K is the number of complex-valued latent components. The detailed derivations of Eqns. (14)-(16) follow from the linear complex Gaussian model.

B. M Step: Update of parameters

The matrix $\boldsymbol{A}_f$ can be found according to

$$\frac{\partial}{\partial \boldsymbol{A}_f} \langle \log p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\theta}) + \log p(\boldsymbol{A}_f) \rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')} = 0$$

$$-\boldsymbol{\Sigma}_{\mathbf{n},f}^{-1} \langle \boldsymbol{R}_{\boldsymbol{xs},f} \rangle + \boldsymbol{\Sigma}_{\mathbf{n},f}^{-1} \boldsymbol{A}_f \langle \boldsymbol{R}_{\boldsymbol{ss},f} \rangle + \boldsymbol{\varphi}(\boldsymbol{A}_f) = 0 \tag{21}$$

where $\boldsymbol{\varphi}(\boldsymbol{A}_f) = \partial \log p(\boldsymbol{A}_f) / \partial \boldsymbol{A}_f$. In the case of $P(\boldsymbol{A}_f)$ is an uniform distribution, then Eqn. (21) leads to a simple closed form expression

$$\boldsymbol{A}_f = \widehat{\boldsymbol{R}}_{\boldsymbol{xs},f} \widehat{\boldsymbol{R}}_{\boldsymbol{ss},f}^{-1}. \tag{22}$$

The matrix $\boldsymbol{\Sigma}_{\mathbf{n},f}$ can be found similarly as

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathbf{n},f}^{-1}} \langle \log p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\theta}) + \log p(\boldsymbol{\Sigma}_{\mathbf{n},f}) \rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')} = 0$$

$$-\boldsymbol{\Sigma}_{\mathbf{n},f} + \boldsymbol{R}_{\boldsymbol{xx},f} - \boldsymbol{A}_f \langle \boldsymbol{R}_{\boldsymbol{xs},f}^{H} \rangle - \langle \boldsymbol{R}_{\boldsymbol{xs},f} \rangle \boldsymbol{A}_f^{H} + \boldsymbol{A}_f \langle \boldsymbol{R}_{\boldsymbol{ss},f} \rangle \boldsymbol{A}_f^{H} + \boldsymbol{\varphi}(\boldsymbol{\Sigma}_{\mathbf{n},f}) = 0 \tag{23}$$

where $\boldsymbol{\varphi}(\boldsymbol{\Sigma}_{\mathbf{n},f}) = \partial \log p(\boldsymbol{\Sigma}_{\mathbf{n},f}) / \partial \boldsymbol{\Sigma}_{\mathbf{n},f}^{-1}$. When $P(\boldsymbol{\Sigma}_{\mathbf{n},f})$ assumes a uniform distribution, then Eqn. (23) leads to

$$\boldsymbol{\Sigma}_{\mathbf{n},f} = diag\left(\boldsymbol{R}_{\boldsymbol{xx},f} - \boldsymbol{A}_f \widehat{\boldsymbol{R}}_{\boldsymbol{xs},f}^{H} - \widehat{\boldsymbol{R}}_{\boldsymbol{xs},f} \boldsymbol{A}_f^{H} + \boldsymbol{A}_f \widehat{\boldsymbol{R}}_{\boldsymbol{ss},f} \boldsymbol{A}_f^{H} \right). \tag{24}$$

Various models exist to model the prior distribution $p(\boldsymbol{A}_f)$ and $p(\boldsymbol{\Sigma}_{n,f})$ which can be incorporated into the above estimation; however, uniform prior distribution results in computational stable and ease of implementation. The determination of $\boldsymbol{W}$ and $\boldsymbol{H}$ will follow the multiplicative update rule. At this point we can distinguish between the two proposed algorithms, and how the targeted speech signal will be described through the exemplar signal.

1) Exemplar Based Algorithm

In this algorithm, the exemplar signal will be used to initialize the targeted speech signal (see Section IV.C) and guide separation through matrix co-partial factorization. The matrix co-partial factorization simultaneously decompose the targeted signal and the side information and force them to partially share the common frequency basis in order to enable the side information to guide the separation of the targeted signal [22,28-29]. In this paper, we propose the NMPCF2D which is a two-dimensional deconvolution of the matrix co-partial factorization. The NMPCF2D uses not only the frequency (spectral) basis but also the convolutive parameters (τ and ϕ) in order to describe the temporal and spectral changes of the targeted speech signal, and therefore renders it more distinguishable and hence more separable than the other sources in the mixture.

The second term in the right hand side of Eqn. (11) can be expressed using the Itakura-Saito divergence with power spectrogram estimated from the E-step. The third term involves the parametrization of $\{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Lambda}\}$. Each element of $\boldsymbol{H}$ has independent decay parameter $\lambda^j_{k,\phi,n}$ with exponential distribution given by $p(\boldsymbol{H}^j|\boldsymbol{\Lambda}^j) = \prod_{k,\phi,n} p\left(h^j_{k,\phi,n}\middle|\lambda^j_{k,\phi,n}\right) = \prod_{k,n,\phi} \lambda^j_{k,\phi,n} \exp\left(-\lambda^j_{k,\phi,n} h^j_{k,\phi,n}\right)$. The prior over $\{\boldsymbol{W}^j\}$ can be assumed flat such that each spectral component is factor-wise normalized to unit length i.e. $p(\boldsymbol{W}^j) = \prod_k \delta\left(\left\|\boldsymbol{W}^j_k\right\|_2 - 1\right)$ where $\left\|\boldsymbol{W}^j_k\right\|_2 = \sqrt{\sum_{f,\tau}\left(w^j_{k,f,\tau}\right)^2}$. Thus, taking the conditional expectation of the negative logarithm of the second and third terms of (11) leads to

$$-\langle \log p(\boldsymbol{C}|\boldsymbol{W}, \boldsymbol{H}) + \log p(\boldsymbol{W}) + \log p(\boldsymbol{H}|\boldsymbol{\Lambda})\rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')}$$
$$= \sum_j \left[\sum_{f,n} D_{IS}\left(\hat{p}_{j,fn} \middle| \sum_{k,\tau,\phi} w^j_{k,f-\phi,\tau} h^j_{k,\phi,n-\tau}\right) - \sum_k \log \delta\left(\left\|\boldsymbol{W}^j_k\right\|_2 - 1\right) + \sum_{k,n,\phi}\left(\lambda^j_{k,\phi,n} h^j_{k,\phi,n} - \log \lambda^j_{k,\phi,n}\right)\right] \quad (25)$$

where $j = \{g, b\}$ and $\hat{p}_{j,fn}$ is the j-th source power spectrogram estimated from (13). Thanks to the E-step, we now have direct access to the estimates of the target speech and background signals in order to estimate $\{\boldsymbol{W}^g, \boldsymbol{W}^b\}$ and $\{\boldsymbol{H}^g, \boldsymbol{H}^b\}$ rather than from the mixture signal which is noisy. We are also able to estimate the mixing gain thanks to the artificial stereophonic channel which augments the dimensionality of the mixing matrix and increases its rank. The separation performance, however, can be weakened under the adverse conditions of low signal-to-interference ratio and the background signal shares some characteristics with the target speech. To alleviate these conditions, a form of exemplar signal whose spectral and temporal

characteristics resemble the target speech will be used. The exemplar signal can be derived from the text associated with the mixture and generated by using a speech synthesizer or human speakers. Let $\tilde{y}(t)$ be the sampled exemplar signal, $y_{f,n_y} \in \mathbb{C}^{1\times N_y}$ be the STFT of $\tilde{y}(t)$, and $p_{y,f,n_y} = |y_{f,n_y}|^2$ is the power spectrogram of the exemplar signal. We want to emphasize that N can differ from N_y due to the temporal mismatch between the exemplar signal and the mixture since it is not practically feasible to emulate the exemplar to be an exact match to the targeted speech signal. These temporal mismatches between the exemplar and the targeted speech signals will result in mismatch between the activation tensors of the exemplar and the targeted speech. A synchronization matrix has been adopted to address this issue [35]. With the exemplar signal, we have developed a joint decomposition using both the mixture and exemplar spectrograms to obtain improved estimates of the spectral basis $\boldsymbol{W}$ and temporal tensors $\boldsymbol{H}$. This is done by allowing the exemplar signal to be factorized using similar model i.e., multi-component NMF2D $p_{y,f,n_y} \approx \sum_{k=1}^{K_g}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w^{y}_{k,f-\phi,\tau}h^{y}_{k,\phi,n-\tau}$. We augment Eqn. (25) with a weighted joint factorization of the exemplar spectrogram as follows

$$\mathcal{J} = \sum_{j,f,n} D_{IS}\left(\hat{p}_{j,fn}\middle|\sum_{k,\tau,\phi} w^{j}_{k,f-\phi,\tau}h^{j}_{k,\phi,n-\tau}\right) + \eta D_{IS}\left(p_{y,fn_y}\middle|\sum_{k,\tau,\phi} w^{y}_{k,f-\phi,\tau}h^{y}_{k,\phi,n_y-\tau}\right) - \sum_{j,k}\log\left(\delta\left(\left\|\boldsymbol{W}^{j}_{k}\right\|_2 - 1\right)\right) + \sum_{j,k,n,\phi}\left(\lambda^{j}_{k,\phi,n}h^{j}_{k,\phi,n} - \log\lambda^{j}_{k,\phi,n}\right)$$

subject to

$$\boldsymbol{W}^{y} = \boldsymbol{W}^{g} \text{ and } \boldsymbol{H}^{y} = \boldsymbol{H}^{g}\boldsymbol{D}^{T}.$$

In above, η is the scalar that weighs the importance of the exemplar signals in the factorization process and $\boldsymbol{D}$ is the synchronization matrix of dimension $N_y \times N$ to ameliorate the temporal mismatch between the exemplar and the mixture. The first two terms on the right hand side represent the matrix factorization of the sources and exemplar spectrograms into the spectral basis and activation tensors, the third term denotes the regularization on the spectral basis, and the fourth tem represents the sparseness of the activation. The regularization involving $\delta\left(\left\|\boldsymbol{W}^{j}_{k}\right\|_2 - 1\right)$ can be satisfied by explicitly normalizing each spectral dictionary to unity i.e. $w^{j}_{k,f,\tau} = w^{j}_{k,f,\tau}/\sqrt{\sum_{f,\tau}\left(w^{j}_{k,f,\tau}\right)^2}$. Using the definition of the Itakura-Saito divergence and by letting $v^{g}_{fn} = \sum_{k,\tau,\phi} w^{g}_{k,f-\phi,\tau}h^{g}_{k,\phi,n-\tau}$, $v^{b}_{fn} = \sum_{k,\tau,\phi} w^{b}_{k,f-\phi,\tau}h^{b}_{k,\phi,n-\tau}$ and $v^{y}_{fn} = \sum_{k,\tau,\phi} w^{y}_{k,f-\phi,\tau}h^{y}_{k,\phi,n-\tau}$, the above cost function reduces up to the constant terms to

$$\mathcal{J} \stackrel{c}{=} \sum_{f,n} \left(\hat{p}_{1,fn}\, v_{fn}^{g^{-1}} - \log\, v_{fn}^{g^{-1}} \right) + \sum_{k,n,\phi} \lambda_{k,\phi,n}^{g} h_{k,\phi,n}^{g} - \sum_{k,n,\phi} \log \lambda_{k,\phi,n}^{g} + \sum_{f,n} \left(\hat{p}_{2,fn}\, v_{fn}^{b^{-1}} - \log\, v_{fn}^{b^{-1}} \right)$$
$$+ \sum_{k,n,\phi} \lambda_{k,\phi,n}^{b} h_{k,\phi,n}^{b} - \sum_{k,n,\phi} \log \lambda_{k,\phi,n}^{b} + \sum_{f,n_y} \eta \left(p_{y,fn_y}\, v_{fn_y}^{y^{-1}} - \log\, v_{fn_y}^{y^{-1}} \right). \tag{26}$$

The multiplicative updates (MU) approach will be used to estimate $\boldsymbol{W}^{g}$ and $\boldsymbol{H}^{g}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \cdot \frac{[\nabla \mathcal{J}]_{-}}{[\nabla \mathcal{J}]_{+}} \tag{27}$$

where $\nabla \mathcal{J} = [\nabla \mathcal{J}]_{+} - [\nabla \mathcal{J}]_{-}$. This leads to

$$w_{k,f,\tau}^{g} \leftarrow w_{k,f,\tau}^{g} \frac{\begin{pmatrix} \sum_{\phi,n} \hat{p}_{1,f+\phi,n} v_{f+\phi,n}^{g^{-2}} h_{k,\phi,n-\tau}^{g} + \\ \eta \sum_{\phi,n_y} p_{y,f+\phi,n_y} v_{f+\phi,n_y}^{y^{-2}} h_{k,\phi,n_y-\tau}^{y} \end{pmatrix}}{\sum_{\phi,n} v_{f+\phi,n}^{g^{-1}} h_{k,\phi,n-\tau}^{g} + \eta \sum_{\phi,n_y} v_{f+\phi,n_y}^{y^{-1}} h_{k,\phi,n_y-\tau}^{y}} \tag{28}$$

given that $\boldsymbol{W}^{y} = \boldsymbol{W}^{g}$, and

$$h_{k,\phi,n}^{g} \leftarrow h_{k,\phi,n}^{g} \left(\frac{\begin{matrix} \sum_{f,\tau} \hat{p}_{1,f,n+\tau} v_{f,n+\tau}^{g^{-2}} w_{k,f-\phi,\tau}^{g} + \\ \eta \sum_{f,\tau,n_y} w_{k,f-\phi,\tau}^{y} p_{y,f,n_y+\tau} v_{f,n_y+\tau}^{y^{-2}} d_{n_y,n} \end{matrix}}{\begin{matrix} \sum_{f,\tau} v_{f,n+\tau}^{g^{-1}} w_{k,f-\phi,\tau}^{g} + \lambda_{k,\phi,n}^{g} + \\ \eta \left(\sum_{f,\tau,n_y} w_{k,f-\phi,\tau}^{y} v_{f,n_y+\tau}^{y^{-1}} d_{n_y,n} + \lambda_{k,\phi,n_y}^{y} d_{n_y,n} \right) \end{matrix}} \right) \tag{29}$$

given that $\boldsymbol{H}^{y} = \boldsymbol{H}^{g}\boldsymbol{D}^{T}$. For the sparsity term, the update is obtained by solving $\frac{\partial}{\partial \lambda_{k,\phi,n}^{g}} \langle \log p(\boldsymbol{C}, \boldsymbol{\theta} | \boldsymbol{X}) \rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')} = 0$ which leads to

$$\lambda_{k,\phi,n}^{g} = \frac{1}{h_{k,\phi,n}^{g}}. \tag{30}$$

By following the same procedure as above, $\boldsymbol{W}^{b}$, $\boldsymbol{H}^{b}$ and $\lambda_{k,\phi,n}^{b}$ can be estimated as follows:

$$w_{k,f,\tau}^{b} \leftarrow w_{k,f,\tau}^{b} \frac{\sum_{\phi,n} \hat{p}_{2,f+\phi,n} v_{f+\phi,n}^{b^{-2}} h_{k,\phi,n-\tau}^{b}}{\sum_{\phi,n} v_{f+\phi,n}^{b^{-1}} h_{k,\phi,n-\tau}^{b}} \tag{31}$$

$$h_{k,\phi,n}^{b} \leftarrow h_{k,\phi,n}^{b} \left(\frac{\sum_{f,\tau} \hat{p}_{2,f,n+\tau} v_{f,n+\tau}^{b^{-2}} w_{k,f-\phi,\tau}^{b}}{\sum_{f,\tau} v_{f,n+\tau}^{b^{-1}} w_{k,f-\phi,\tau}^{b} + \lambda_{k,\phi,n}^{b}} \right) \tag{32}$$

$$\lambda_{k,\phi,n}^{b} = \frac{1}{h_{k,\phi,n}^{b}} \tag{33}$$

Similarly, $\boldsymbol{W}^{y}$ and $\boldsymbol{H}^{y}$ can be estimated as

$$w_{k,f,\tau}^{y} \leftarrow w_{k,f,\tau}^{y} \frac{\sum_{\phi,n_y} p_{y,f+\phi,n_y} v_{f+\phi,n_y}^{y^{-2}} h_{k,\phi,n_y-\tau}^{y}}{\sum_{\phi,n_y} v_{f+\phi,n_y}^{y^{-1}} h_{k,\phi,n_y-\tau}^{y}} \tag{34}$$

$$h_{k,\phi,n_y}^{y} \leftarrow h_{k,\phi,n_y}^{y} \left(\frac{\sum_{f,\tau} p_{y,f,n_y+\tau} v_{f,n_y+\tau}^{y^{-2}} w_{k,f-\phi,\tau}^{y}}{\sum_{f,\tau} v_{f,n_y+\tau}^{y^{-1}} w_{k,f-\phi,\tau}^{y}} \right). \tag{35}$$

2) Semi-Exemplar Based Algorithm

A variant of the above algorithm is to use the exemplar signal only for initializing the target speech signal. The MU rule will be used to update $\{\boldsymbol{W}^j, \boldsymbol{H}^j\}_{j=\{g,b\}}$. These are obtained by setting $\eta = 0$ for Eqns. (28)-(29) leading to

$$w_{k,f,\tau}^{g} \leftarrow w_{k,f,\tau}^{g} \frac{\sum_{\phi,n} \hat{p}_{1,f+\phi,n} v_{f+\phi,n}^{g^{-2}} h_{k,\phi,n-\tau}^{g}}{\sum_{\phi,n} v_{f+\phi,n}^{g^{-1}} h_{k,\phi,n-\tau}^{g}} \tag{36}$$

$$h_{k,\phi,n}^{g} \leftarrow h_{k,\phi,n}^{g} \left(\frac{\sum_{f,\tau} \hat{p}_{1,f,n+\tau} v_{f,n+\tau}^{g^{-2}} w_{k,f-\phi,\tau}^{g}}{\sum_{f,\tau} v_{f,n+\tau}^{g^{-1}} w_{k,f-\phi,\tau}^{g} + \lambda_{k,\phi,n}^{g}} \right) \tag{37}$$

and the same sparsity in Eqn. (30) will be used. The tensors of the background signal follow similarly and they are shown in Eqns. (31)-(32), while the sparsity update in Eqn. (33).

The semi-exemplar based algorithm uses the exemplar signal to initialize the tensors of the NMPCF2D and, thus it depends on the exemplar to give a good start. On the other hand, the exemplar based algorithm proposed in Section III B 1) uses the exemplar signal not only to give the correct initialization but also to guide the whole algorithm through the NMPCF2D which jointly factorizes the exemplar and mixture signals to yield the desired spectral bases and temporal code. Therefore, the exemplar based algorithm recycles the signal $\tilde{y}(t)$ more often than the semi-exemplar based algorithm. However, the latter does not require the synchronization matrix $\boldsymbol{D}$ which is time-consuming. In addition, the temporal activation is constraint-free to adapt according to the underlying mixture signal.

IV. Describing Target Speech using Exemplar Signal

The description of the target speech signal will be carried out indirectly by the exemplar signal and from the aid of the NMF2D that optimizes its parameters. Here the parameters of the NMF2D will be optimized by depending on the exemplar signal instead of the mixture. The exemplar is considered instead of the targeted speech signal as it is unavailable. The NMF2D is proposed due to its ability in descripting the temporal and spectral changes through the convolutive parameters (τ and ϕ), and specifying the required number of frequency basis K.

A. *Components Order Selection*

The determination of the model order for the NMF2D will be realized by using the exemplar signal $y(t)$:

Step 1: Optimize W^y, and H^y by using (34) and (35).

Step 2: Optimizing τ and ϕ

Set $K = 1$

For $\tau_{max} = 1$ to T

For $\phi_{max} = 1$ to Φ

- Estimate $v_{fn}^y = \sum_{k,\tau,\phi} w_{k,f-\phi,\tau}^y h_{k,\phi,n-\tau}^y$.
- Estimate the signal-to-distortion ratio (SDR) [36] between the exemplar signal p_{y,fn_y} and its approximation v_{fn}^y in order to evaluate the factorization performance.

Select convolutive parameters (τ_{max}, ϕ_{max}) with the highest SDR.

Step 3: Optimizing K

For $K = 2$ to K_{max}

- Estimate $v_{fn}^y = \sum_{k,\tau,\phi} w_{k,f-\phi,\tau}^y h_{k,\phi,n-\tau}^y$.
- Estimate the SDR between the exemplar signal p_{y,fn_y} and its approximation v_{fn}^y.

Select K with the highest SDR.

B. *Components Reconstruction*

The estimated sources $\hat{\boldsymbol{s}}_{fn}$ can be reconstructed by using Wiener filtering $\left(\boldsymbol{\Sigma}_{\boldsymbol{s},fn}\boldsymbol{A}_f^H\boldsymbol{\Sigma}_{\boldsymbol{x},fn}^{-1}\right)$ as in Eqn. (14), and due to the linearity of the STFT, the inverse-STFT (with dual synthesis window [37]) can be used to transform it to the time domain.

C. *Initialization*

The initialization is an essential part for the separation since the NMF2D is very sensitive to the initialization. In this paper, we initialize the spectral and temporal tensors for the proposed algorithms using the exemplar signal $\tilde{y}(t)$ which itself is decomposed into $w_{k,f,\tau}^y$ and $h_{k,\phi,n}^y$:

$$\left(w_{k,f,\tau}^g\right)_{ini} = w_{k,f,\tau}^y$$

$$\left(h_{k,\phi,n}^g\right)_{ini} = h_{k,\phi,n}^y d_{n_y,n}$$

where $d_{n_y,n}$ is synchronization parameter. For the background, $\left(w_{k,f,\tau}^b\right)_{ini}$ and $\left(h_{k,\phi,n}^b\right)_{ini}$ will be randomly initialized. Thus we can initialize the spectral bases and temporal activation for the mixture as follows:

$$\left(w_{k,f,\tau}^{x}\right)_{ini} = \left[\left(w_{k,f,\tau}^{g}\right)_{ini} \quad \left(w_{k,f,\tau}^{b}\right)_{ini}\right]$$

$$\left(h_{k,\phi,n}^{x}\right)_{ini} = \begin{bmatrix}\left(h_{k,\phi,n}^{g}\right)_{ini} \\ \left(h_{k,\phi,n}^{b}\right)_{ini}\end{bmatrix}.$$

Table I summarizes the proposed full-exemplar based algorithm. In the case of semi-exemplar based algorithm, the M-step will be performed differently. In particular, $\left\{w_{k,f,\tau}^{g}, h_{k,\phi,n}^{g}\right\}$ are computed by (36)-(37).

V. Results and Discussions

A. Dataset

The performance of the proposed algorithms will be investigated and compared with recent text and music informed source separation methods [22]. For fair comparison, the same datasets will be used. The dataset is derived from 10 speech mixtures that mix with music (Speech + music) and effect (Speech + Fx). For each mixture the speech is emulated by using 12 exemplars (synth Man, Synth Woman, TMT Man, TMT woman, and other 8 foreign speakers). Thus we will have 240 experiments (generated from the 20 mixtures and the 12 exemplars for each mixture) for SNR of -5dB.

B. Evaluation

In order to evaluate the proposed algorithms, SDR is adopted which combines the source-to-interference ratio (SIR) and the source-to-artefacts ratio (SAR). The MATLAB codes for this evaluation procedure can be found in [38].

C. Selections of $\eta, \delta,$ and γ

The contribution of the exemplar on the separation is weighted by η, so if $\eta \to 0$ the exemplar will have little effect, while if its value is increased the exemplar will have more influence. To determine the value of η we use the following:

$$\eta = \eta_0 \frac{N}{N_y} \tag{38}$$

where N and N_y is the temporal length of the mixture and the exemplar, respectively, and for our case we consider $\eta_0 = 0.5$. The factor $e^{-i\omega\delta}$ that appears in the artificial stereophonic channel (in Eqn. (6)) is only uniquely specified if $|\omega\delta| < \pi$, otherwise this would cause phase-wrap. Selecting improper time-delay δ will lead to phase-wrap if the maximum frequency of the source is exceeded. In order to avoid phase ambiguity, we must satisfy

$$|\omega_{max}\delta_{max}| < \pi. \tag{39}$$

where $\omega_{max} = 2\pi f_{max}/f_s$, δ_{max} is the maximum time delay, f_{max} is the maximum frequency present in the sources and f_s is the sampling frequency. Hence, δ_{max} can be determined from Eqn. (39) according to

$$\delta_{max} < \frac{f_s}{2f_{max}} \tag{40}$$

As long as the delay parameter is less than δ_{max}, there will not be any phase ambiguity. Assuming a maximum frequency $f_{max} = 3.5$ kHz, and a sampling frequency $f_s = 16$ kHz, one obtains $\delta_{max} < 2.3$ using Eqn. (40). Therefore, phase ambiguity can be avoided when δ is selected up to 2.3. Additionally, for a maximum frequency $f_{max} = 8$ kHz the maximum delay δ_{max} is limited to 1. This condition is used to determine the range of δ in formulating the artificial stereophonic channel. For the weighting parameter γ that determines the attenuation on the delayed mixture $\gamma\tilde{x}_1(t-\delta)$ (see Eqn. (2)), we found that exists a range of γ with high SDR as shown in Fig. 2. The plot suggests that this range to be $0.1 \leq \gamma \leq 0.25$. In all our cases, we use $\gamma = 0.15$.

D. OPTIMIZATION OF $\tau, \phi,$ and K

By following the procedure described in Section IV A (i.e. setting $\mathrm{T} = 10$, $\Phi = 10$, and $K_{max} = 10$), the results for one exemplar are shown in Fig. 3. Fig. 3(a) shows that the best SDR is attained at $\tau = 9$ and $\phi = 1$. In addition, Fig. 3(b) reveals that $K = 2$ results in the optimum number of components. The parameters of the exemplars for all 10 mixtures are computed resulting into 120 different parameters ($\tau, \phi,$ and K). This number is due to 120 different exemplars (each speech signal in the mixture is emulated by 12 exemplars, and as there are 10 mixtures, this results in 120 exemplars)[†]. Despite 12 exemplars emulating the same speech signal, they have different parameters because they derived from different speakers (native and non-native English) and different genders, and as a result of these differences we have different parameters of the NMF2D that describe each exemplar.

E. RESULTS

The STFT windows length is set to 512 with 50% overlaps. To show the convergence of the proposed algorithms, the convergence of the cost functions Eqn. (11) of both algorithms are shown in Fig. 4. This plot is obtained for one mixture with twelve exemplars. It is noted that all trajectories have converged to the steady state in less than 50 iterations. The fast and stable convergence is attributed to the manner of how the GEM-MU algorithm adapts the model parameters and latent variables.

[†] Both 120 (Speech+Music) mixture group and 120 (Speech+Effects) mixture group have the same speech signal.

1) Effects of the exemplar on the separation: In this subsection, the effects of the exemplar on the separation performance will be studied. Four cases will be considered in the following depending on how the exemplar is constructed with respect to the targeted speech signal.

Case 1: Exemplar is identical to the targeted speech signal

This is considered as an ideal case since it is not always possible to emulate exactly the targeted speech signal; however, it forms the baseline for further comparisons. This case is considered by taking the targeted speech source as an exemplar. The results are tabulated in Table II. Since the exemplar is identical to the targeted speech signal, it gives a good initialization for the semi-exemplar based algorithm. In addition, the (full-)exemplar based algorithm benefitted from both good initialization and temporal tracking of the targeted speech signal during the whole time. The waveforms of the mixture, the original speech signal, and its corresponding estimate are shown in Fig. 5. It is clear that both algorithms have correctly estimated the source.

Case 2: Variation of exemplar's gender

Depending on the gender of the exemplar two cases will be considered. First, the same gender case where the exemplar comes from the same gender as the targeted speech signal (either both males or both females). Second, the different gender case where the exemplar is derived from different gender from the targeted speech signal (one male and the other is female, and vice versa). The obtained separation results are tabulated in Table III. The results show that similar gender exemplar gives higher SDR performance than the different gender. This difference is acceptable as the male cannot emulate effectively the female voice and vice versa, therefore a difference occurs between the two cases. Furthermore, the waveforms of the speech signal and their estimates are shown in Fig. 6. The plot shows that in both cases the proposed algorithm has estimated the speech signal correctly thanks to the NMPCF2D in removing the temporal and spectral mismatches between the exemplar and the targeted speech signal. However, the same gender case results in slightly better SDR as the exemplar is more similar to the targeted speech signal than the different gender exemplar.

Case 3: Variation of exemplar's native speaker

Depending on the exemplar whether it is native English speaker or not, two cases are correspondingly considered. The first case corresponds to both targeted speech signal and exemplar being native English speaker. The second case corresponds to the situation where the targeted speech signal is native English speaker while the exemplar is non-native English speaker. In the latter, it is difficult for the non-native English speaker to emulate the targeted speech signal in the same way as the native English speaker, due to the different phoneme expression and accent of the two speakers, as shown in the waveforms of the native and

non-native exemplar in Fig. 7. The plot shows that the non-native English speaker waveform is substantially different than the targeted speech signal in comparison with the native English speaker signal. The difference has clearly influenced the accuracy of the estimated signal as shown in the plot. The results are tabulated in Table III which indicates that the similar native exemplar tends to yield better separation than the non-native exemplar. It is interesting to note that in comparison with Case 2, the effects of native/non-native speakers tend to have stronger impact on the separation performance than that of gender differences. An average reduction of 0.44dB is observed when comparing the SDR performance between Case 2 and Case 3.

Case 4: Missing information exemplar

In this case, the exemplar does not emulate all the words in the targeted speech signal. This case has been achieved by removing 20% of information in the exemplar. The results are tabulated in Table III, which indicates a difference between using the complete and missing words exemplar. The result shows the non-trivial effects of the exemplar on the separation algorithm as it guides the factorization process for the whole time series. Despite the exemplar is ambiguous due to the missing words, the NMPCF2D has to an extent successfully reconstructed the waveform of targeted speech signal. Fig. 8 shows the obtained results where it is noticeable that the first half of both estimated waveforms are quite similar while the second half shows some aspect of visual differences caused by the missing information exemplar to underestimate the part of the targeted speech signal at those time period when the exemplar is muted.

The above cases demonstrate the influences the exemplar signal has on the separation performance. The more similar the exemplar to the targeted speech signal, the better will be the performance, and vice versa. However, we have also examined the deviation of the exemplar signal from the targeted speech and the obtained results have unanimously indicated that the proposed exemplar algorithm has been able to maintain a relatively robust separation performance.

2) Comparison with recently developed methods: The proposed algorithms will be compared with the matrix factorization model based on the excitation-filter channel speech model [22]. In this algorithm, the variations between the speech example and the targeted speech in the mixture such as pitch variation, pronounced phonemes, recording conditions, and speaker's vocal tract are modelled by the excitation-filter channel speech model. The excitation-filter channel model jointly factorizes the spectrograms of the mixture and the exemplar that emulate the speech signal. Also, we compared with the Structural Gaussian Scaled Mixture Model (GSMM) [22] with constraints applied on the matrices of the excitation-filter channel speech model in order to have a physical motivation, such as allowing one phoneme to be pronounced at a time and one fundamental

frequency to be active at a time. The proposed algorithms have also been compared with Schmidt's algorithm [31] which is based on the conventional NMF2D.

The SDRs of the informed excitation-filter channel speech model, structural GSMM, Schmidt's algorithm, and the proposed algorithms are tabulated in Table IV. The table indicates that the proposed algorithms have better performance than the informed excitation-filter channel speech model, which can be summarized as follows: An achievement of 2.57 dB more for the speech and music group, and 1.89 dB more for the speech and effects group for the semi-exemplar based algorithm. For the exemplar-based algorithm an achievement of 3.12 dB more for the speech and music group, and 3.37 dB more for the speech and effects group. Furthermore, the exemplar based algorithm achieved an improvement of 1.86 dB for the speech and effects group and 0.16 dB for the speech and music group, in comparison with GSMM algorithm. On the other hand, the semi-exemplar based algorithm achieves 0.38 dB more for the speech and effects group, and 0.39 dB less for the speech and music group. Although the proposed semi-exemplar based algorithm is less dependent on the exemplar signal, its high performance is attributed to the artificial stereophonic channels and leverages on the diversity of the full rank mixing matrix. Furthermore, an achievement of 0.65 dB more, and 1.22 dB more for the semi-exemplar based algorithm in comparison with Schmidt's algorithm for both the speech and music group and the speech and effects group, respectively. Finally an achievement of 1.2 dB more, and 2.70 dB more for the exemplar-based algorithm in comparison with Schmidt's algorithm for both the speech and music group, and the speech and effects group, respectively.

In addition to the above, we will examine the source representation of the algorithms. To show the effects of source representation, one component of $\boldsymbol{W}$ and $\boldsymbol{H}$ tensors and their corresponding product for both the GSMM and the proposed NMPCF2D have been plotted in Figs. 9(a) and 9(b), respectively. Both plots show how $\boldsymbol{W}$ models the changes in the frequencies of the source and $\boldsymbol{H}$ the distribution in the time domain. On the separate hand, while $\boldsymbol{W}$ and $\boldsymbol{H}$ of the GSMM detected the frequency bases, they were not able to address the frequency and the temporal changes. Additionally, the spectrogram of the original speech, the exemplar, the mixture, and the estimated speech by using the proposed algorithms and the structural GSMM are shown in Fig. 10. These plots clearly show that the proposed algorithms have successfully detected the pitch and temporal change of the source due to its two-dimensional deconvolution while the structural GSMM failed to detect these changes. Furthermore, Fig. 11 shows the waveforms of these signals. Finally, from Table IV it can be seen that the exemplar based algorithm has achieved better separation results than the semi-exemplar based algorithm since the latter only uses the exemplar to initialize the tensors of the targeted speech signal. Thus the initialization will guide the algorithm for the first iteration and gives the correct start but it may get trapped in local minima or drifted away from the solution as the iterations increases. Although the exemplar based algorithm has been given the identical start as the semi-exemplar based algorithm, its separation is guided by

the NMPCF2D which models both exemplar and the targeted speech signal. To show this, the waveform of the original voice, exemplar, and the estimated voice by using these two algorithms are shown in Fig. 12. The plot indicates that the exemplar based algorithm has successfully estimated the original source. This shows the importance and contribution of NMPCF2D on the proposed algorithm.

VI. CONCLUSION

In this paper two algorithms for the informed single-channel source separation i.e., the semi-exemplar based algorithm and the exemplar-based algorithm, have been proposed. These algorithms leverage on the two-dimensional matrix factorization method, namely, the NMF2D and the proposed NMPCF2D. These algorithms have the advantage of describing the target signal by describing the pitch and temporal changes of that signal, which cannot carry out by the NMF or NMPCF. Artificial stereophonic channel is introduced to render the ill-posed single-channel source separation into an exact-determined system. For faster convergence and better performance, the parameters of both algorithms are adapted using the hybrid GEM-MU algorithm with adaptive sparsity. It has been shown that the proposed method outperformed the conventional algorithms.

REFERENCES

1. Hu D, Xu J, Yu X. Blind Source Separation: Theory and Applications. JohnnWiley & Sons; 2014. 416 p.
2. Mitianoudis N, Davies ME. Audio source separation: Solutions and problems. International Journal of Adaptive Control and Signal Processing. 2004; 18(6):299–314.
3. Comon P, Jutten C. Handbook of Blind Source Separation: Independent component analysis and applications: Academic Press, 2010.
4. Zha D, Qiu T. A new blind source separation method based on fractional lower-order statistics. International Journal of Adaptive Control and Signal Processing. 2006; 20(5): 213–223.
5. Andrzej C, Zdunek R, Phan AH, and Amari S. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Wiley & Sons; 2009. 500 p.
6. Moir TJ, Harris JI. Decorrelation of multiple non-stationary sources using a multivariable crosstalk-resistant adaptive noise canceller. International Journal of Adaptive Control and Signal Processing. 2013; 27(5): 349–367.
7. Djendi M. A new two-microphone Gauss-Seidel pseudo affine projection algorithm for speech quality enhancement. International Journal of Adaptive Control and Signal Processing. 2017 (online)

8. He X, He F, Zhu T. Large-scale super-Gaussian sources separation using Fast-ICA with rational nonlinearities. International Journal of Adaptive Control and Signal Processing. 2017; 31(3): 379–397.
9. Fritsch J, Plumbley MD. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013; 888-891.
10. Hennequin R, Bertrand D, Badeau R. Score informed audio source separation using a parametric model of non-negative spectrogram. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011; 45-48.
11. Ewert S, Pardo B, Muller M, Plumbley MD. Score-informed source separation for musical audio recordings: An overview. IEEE Signal Processing Magazine. 2014, 31(3): 116-124.
12. Ozerov A, Fevotte C, Blouet R, Durrieu JL. Multichannel Nonnegative Tensor Factorization with Structured Constraints for User-Guided Audio Source Separation. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011; 257-260.
13. Duong NQK, Ozerov A, Chevallier L. Temporal annotation-based audio source separation using weighted nonnegative matrix factorization. Proc. IEEE 4th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), 2014; 220-224.
14. FitzGerald D. User assisted separation using tensor factorisations. Proc. 20th European Signal Processing Conference (EUSIPCO), 2012; 2412-2416.
15. Smaragdis P. User guided audio selection from complex sound mixtures. Proc. 22nd Annual ACM Symposium on User Interface Software and Technology, 2009; 89-92.
16. Hennequin R, Burred J, Maller S, Leveau P. Speech-guided source separation using a pitch-adaptive guide signal model," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014; 6672 - 6676.
17. Gerber T, Dutasta M, Girin L, Févotte C. Professionally-produced music separation guided by covers. Proc. International Society for Music Information Retrieval Conference (ISMIR 2012), Porto, Portugal, 2012; 85-90.
18. Souviraa-Labastie N, Vincent E, Bimbot F. Music separation guided by cover tracks: Designing the joint NMF model. IEEE Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015; 484-488.
19. Souviraa-Labastie N, Olivero A, Vincent E, Bimbot F. Multi-channel audio source separation using multiple deformed references. IEEE/ACM Transactions on Audio Speech and Language Processing, 2015; 23(11): 1775-1787.

20. Souviraa-Labastie N, Olivero A, Vincent E, Bimbot F. Audio source separation using multiple deformed references. Proc. of European Signal Processing Conference (EUSIPCO), 2014; 311-315.
21. Liutkus A, Leveau P. Separation of Music+Effects sound track from several international versions of the same movie. Proc. of AES 128th Convention, London, United Kingdom, 2010.
22. Magoarou LL, Ozerov A, Duong NQK. Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization. Journal of Signal Processing Systems for Signal Image and Video Technology, 2015; 79(2): 117-131.
23. Zhang X, Su CY, Lin Y, Ma L, Wang J. Adaptive neural network dynamic surface control for a class of time-delay nonlinear systems with hysteresis inputs and dynamic uncertainties. IEEE Transactions on Neural Networks and Learning Systems. 2015; 26: 2844-2860.
24. Zhang X, Li Z, Su CY, Lin Y, Fu Y. Implementable adaptive inverse control of hysteretic systems via output feedback with application to piezoelectric positioning stages. IEEE Transactions on Industrial Electronics. 2016; 63: 5733-5743.
25. Zhang X, Xu Z, Su CY, Li Z, Li X, Xiong C, Lin Y. Fuzzy approximator based adaptive dynamic surface control for unknown time delay nonlinear systems with input asymmetric hysteresis nonlinearities. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2017; 47: 2218-2232.
26. Liutkus A, Pinel J, Badeau R, Girin L, Richard G. Informed source separation through spectrogram coding and data embedding. Signal Processing, 2012; 92(8): 1937-1949.
27. Ozerov A, Liutkus A, Badeau R, Richard G. Informed source separation: Source coding meets source separation. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011; 257-260.
28. Liutkus A, Gorlow S, Sturmel N, Zhang SH, Girin L, Badeau R, Daudet L, Marchand S, Richard G. Informed audio source separation: A comparative study. Proc. of 20th European Signal Processing Conference (EUSIPCO), 2012; 2397-2401.
29. Liutkus A, Ozerov A, Badeau R, Richard G. Spatial coding-based informed source separation. Proc. of 20th European Signal Processing Conference (EUSIPCO), 2012; 2407-2411.
30. Ozerov A, Liutkus A, Badeau R, Richard G. Coding-based informed source separation: Nonnegative tensor factorization approach. IEEE Transactions on Audio Speech and Language Processing, 2013, 21(8): 1699-1712.
31. M. N. Schmidt, and M. Morup, “Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,” in 6th Intl. Conf. on Independent Component Analysis and Signal Separation (ICA ’06), Charleston, USA, 2006, pp. 700–707.

32. Yoo J, Kim M, Kang K, Choi S. Nonnegative matrix partial co-factorization for drum source separation. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010; 1942-1945.
33. Kim M, Yoo J, Kang K, Choi S. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. IEEE Journal of Selected Topics in Signal Processing, 2011; 5(6): 1192-1204.
34. Neeser FD, Massey JL. Proper complex random-processes with applications to information-theory. IEEE Trans. on Information Theory, 1993; 39(4): 1293-1302.
35. Pedone A, Burred JJ, Maller S, Leveau P. Phoneme-level text to audiosynchronization on speech signals with background music. Proc. of 12th Conference of the International Speech Communication Association, 2011; 433–436.
36. Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation. IEEE Trans. on Audio Speech and Language Processing, 2006; 14(4): 1462-1469.
37. Goodwin M. The STFT, sinusoidal models, and speech modification. Springer Handbook of Speech Processing, J. Benesty, M. M. Sondhi and Y. Huang, eds., New York: Springer, 2008.
38. Signal separation evaluation campaign (SiSEC 2018). http://sisec.wiki.irisa.fr.

APPENDIX

The separability of model can be examined from the pseudo-stereo mixture by considering $a_j(\delta)$ in the three cases. Case 1 refers to identical sources mixed in the single channel, Case 2 represents different sources but setting γ and δ for the pseudo-stereo mixture such that $a_{\tilde{g}}(t;\delta,\gamma) = a_{\tilde{b}}(t;\delta,\gamma)$, and Case 3 corresponds to the most general case where the sources are distinct, and γ and δ are selected arbitrarily such that the mixing attenuations and residues are also different. The above cases are evaluated in the light of the following sepability function:

$$J(n,f) = \underset{k\in(\tilde{g},\tilde{b})}{\operatorname{argmin}} \left| \bar{a}_k(n,f) e^{-i\omega\delta} x_{1,f,n} - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) x_{1,f,n} \right|^2 \tag{A1}$$

where

$$\bar{a}_{\tilde{g}}(n,f) = a_{\tilde{g}}(\delta) - C_{\tilde{g}}(n,f)$$

$$\bar{a}_{\tilde{b}}(n,f) = a_{\tilde{b}}(\delta) - C_{\tilde{b}}(n,f)$$

$$C_{\tilde{g}}(n,f) = \frac{1}{1+|\gamma|} \sum_{\substack{m=1 \\ m\neq\delta}}^{D_k} a_{\tilde{g}}(m) e^{-i\omega(m-\delta)}$$

$$C_{\tilde{b}}(n,f) = \frac{1}{1+|\gamma|} \sum_{\substack{m=1 \\ m\neq\delta}}^{D_k} a_{\tilde{b}}(m) e^{-i\omega(m-\delta)}$$

Technically, this function partitions the TF plane of the mixed signal into k groups of (n, f) units by evaluating the cost function. For each TF unit, the k^{th} argument that gives the minimum cost will be assigned to the k^{th} source. We may analyze further by assuming that the target source dominates at a particular TF unit. In this case, the first line of (15) reduces to $x_{1,f,n} = g_{f,n}$ and therefore, the above becomes

$$J(n,f) = \underset{k}{\operatorname{argmin}} \left| \bar{a}_k(n,f) e^{-i\omega\delta} g_{f,n} - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) g_{f,n} \right|^2$$

$$= \underset{k}{\operatorname{argmin}} \left| \bar{a}_k(n,f) e^{-i\omega\delta} g_{f,n} - \frac{g_{f,n}}{1+|\gamma|} - \frac{\gamma e^{-i\omega\delta}}{1+|\gamma|} g_{f,n} \right|^2$$

$$= \underset{k}{\operatorname{argmin}} \left| \bar{a}_k(n,f) e^{-i\omega\delta} g_{f,n} + \sum_{m=1}^{D_{\tilde{g}}} \frac{a_{\tilde{g}}(m) e^{-i\omega m}}{1+|\gamma|} g_{f,n-m} - \frac{\gamma e^{-i\omega\delta}}{1+|\gamma|} g_{f,n} \right|^2$$

$$= \underset{k}{\operatorname{argmin}} \left| \bar{a}_k(n,f) e^{-i\omega\delta} g_{f,n} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_{\tilde{g}}} \frac{a_{\tilde{g}}(m) e^{-i\omega m}}{1+|\gamma|} g_{f,n-m} - \left(\frac{-a_{\tilde{g}}(\delta)+\gamma}{1+|\gamma|}\right) e^{-i\omega\delta} g_{f,n} \right|^2$$

$$= \underset{k}{\operatorname{argmin}} \left| a_k(\delta) e^{-i\omega\delta} g_{f,n} - C_k(n,f) e^{-i\omega\delta} g_{f,n} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_{\tilde{g}}} \frac{a_{\tilde{g}}(m) e^{-i\omega m}}{1+|\gamma|} g_{f,n-m} - a_{\tilde{g}}(\delta) e^{-i\omega\delta} g_{f,n} \right|^2 \quad \text{(A2)}$$

The following three cases are considered:

Case 1: If $a_{\tilde{b}}(\delta) = a_{\tilde{g}}(\delta) = a(\delta)$ and $r_{\tilde{g}}(t) = r_{\tilde{b}}(t) = r(t)$, then $\tilde{x}_2(t) = \left(\frac{a(\delta)+\gamma}{1+|\gamma|}\right) \tilde{x}_1(t-\delta) + 2r(t)$.

In this case, there is no benefit achieved at all. The second mixture is simply formulated as a time-delayed of the first mixture multiply by a scalar plus the redundant residue. The separability of this case is presented by substituting the pseudo-stereo mixture of Case 1 into the cost function. Since both residues are equal, then $C_{\tilde{g}}(n,f) = C_{\tilde{b}}(n,f) = C(n,f) = \frac{1}{1+|\gamma|} \sum_{\substack{m=1 \\ m\neq\delta}}^{D} a(m) e^{-i\omega(m-\delta)}$. For Case 1, the cost function becomes:

$$J(n,f) = \underset{k}{\operatorname{argmin}} \left| a(\delta) e^{-i\omega\delta} g_{f,n} - C(n,f) e^{-i\omega\delta} g_{f,n} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D} \frac{a(m) e^{-i\omega m}}{1+|\gamma|} g_{f,n-m} - a(\delta) e^{-i\omega\delta} g_{f,n} \right|^2 \quad \text{(A3)}$$

Invoking the local stationarity of the source $g_{f,n-D} = g_{f,n}$ for a constant D, then (A3) leads to

$$J(n,f) = \underset{k}{\operatorname{argmin}} \left| \sum_{\substack{m=1 \\ m\neq\delta}}^{D} \frac{\left(a(m) e^{-i\omega m} - a(m) e^{-i\omega m}\right)}{1+|\gamma|} \right|^2 |g_{f,n}|^2 \quad \text{(A4)}$$

$$= 0 \quad \text{for } \forall k.$$

As a result, the cost function $J(n,f)$ is zero for all k arguments. In this case, the cost function cannot distinguish the k arguments, the mixture is not separable.

Case 2: If $a_{\tilde{b}}(\delta) = a_{\tilde{g}}(\delta) = a(\delta)$ and $r_{\tilde{b}}(t) \neq r_{\tilde{g}}(t)$, then $\tilde{x}_2(t) = \left(\frac{a(\delta)+\gamma}{1+|\gamma|}\right) \tilde{x}_1(t-\delta) + r_{\tilde{g}}(t) + r_{\tilde{b}}(t)$.

This case remains almost similar to the previous case and differs only in terms of $r_{\tilde{g}}(t) \neq r_{\tilde{b}}(t)$. As each

residue $r_j(t)$ is related to the j^{th} source via $C_j(n,f)$, the identifiability of this mixture can be analyzed as

$$J(n,f) = \underset{k}{\operatorname{argmin}} \left| a(\delta)e^{-i\omega\delta} g_{f,n} - C_k(n,f)e^{-i\omega\delta} g_{f,n} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_{\tilde{g}}} \frac{a_{\tilde{g}}(m)e^{-i\omega m}}{1+|\gamma|} g_{f,n-m} - a(\delta)e^{-i\omega\delta} g_{f,n-m} \right|^2$$

$$= \underset{k}{\operatorname{argmin}} \left| \sum_{\substack{m=1 \\ m\neq\delta}}^{D_{\tilde{g}}} \frac{\left(a_{\tilde{g}}(m) - a_k(m)\right)}{1+|\gamma|} e^{-i\omega m} \right|^2 \left| S_j(\tau,\omega) \right|^2 \quad \text{(A5)}$$

It can be deduced from above that the cost function yields a zero value for $k = \tilde{g}$ i.e. corresponds to the target source, and nonzero value for $k \neq \tilde{g}$. Despite the mixing attenuation for both sources are identical, the cost function is still able to distinguish the k arguments by using only the difference of residues. Therefore, the mixture of Case 2 is separable.

Case 3: If $a_{\tilde{b}}(\delta) \neq a_{\tilde{g}}(\delta)$ and $r_{\tilde{b}}(t) \neq r_{\tilde{g}}(t)$, (or $r_{\tilde{b}}(t) = r_{\tilde{g}}(t)$) then $\tilde{x}_2(t) = \left(\frac{a_{\tilde{g}}(\delta)+\gamma}{1+|\gamma|}\right)\tilde{g}(t-\delta) + \left(\frac{a_{\tilde{b}}(\delta)+\gamma}{1+|\gamma|}\right)\tilde{b}(t-\delta) + r_{\tilde{g}}(t) + r_{\tilde{b}}(t)$.

We first treat the situation of $r_{\tilde{g}}(t) = r_{\tilde{b}}(t)$. Since the mixing attenuations $a_{\tilde{g}}(\delta)$ and $a_{\tilde{b}}(\delta)$ correspond respectively to $\tilde{g}(t)$ and $\tilde{b}(t)$ then the cost function can be expressed as

$$J(n,f) = \underset{k}{\operatorname{argmin}} \left| a_k(\delta)e^{-i\omega\delta} g_{f,n} - C(n,f)e^{-i\omega\delta} g_{f,n} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D} \frac{a(m)e^{-i\omega m}}{1+|\gamma|} g_{f,n-m} - a_{\tilde{g}}(\delta)e^{-i\omega\delta} g_{f,n} \right|^2$$

$$= \underset{k}{\operatorname{argmin}} \left| \left(a_k(\delta) - a_{\tilde{g}}(\delta)\right)e^{-i\omega\delta} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D} \frac{\left(a(m) - a(m)\right)}{1+|\gamma|} e^{-i\omega m} \right|^2 \left| g_{f,n} \right|^2 \quad \text{(A6)}$$

$$= \underset{k}{\operatorname{argmin}} \left| \left(a_k(\delta) - a_{\tilde{g}}(\delta)\right)e^{-i\omega\delta} \right|^2 \left| g_{f,n} \right|^2$$

This cost function yields a nonzero value only for $k \neq \tilde{g}$ i.e. does not correspond to the target source. In this case, the cost function can separate the k arguments due to the difference of a_k and $a_{\tilde{g}}$. The case of $r_{\tilde{g}}(t) \neq r_{\tilde{b}}(t)$ follows similar line of argument as above where the cost function becomes

$$J(n,f) = \underset{k}{\operatorname{argmin}} \left[\left| \left(a_k(\delta) - a_{\tilde{g}}(\delta)\right)e^{-i\omega\delta} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \frac{\left(a_{\tilde{g}}(m) - a_k(m)\right)}{1+|\gamma|} e^{-i\omega m} \right|^2 \left| g_{f,n} \right|^2 \right] \quad \text{(A7)}$$

This cost function yields a nonzero value only for $k \neq \tilde{g}$; thus the cost function is able to distinguish the k arguments. In summary, by considering $a_{\tilde{g}}(t)$ and $r_{\tilde{g}}(t)$ with respect to above three cases, only Case 2 and Case 3 are separable.

Table I

Proposed algorithm

Proposed algorithm
1. Optimize the convolutive parameters and number of components based on the exemplar.
2. Initialize $w_{k,f,\tau}^{g}$and $h_{k,\phi,n}^{g}$ based on the exemplar, $w_{k,f,\tau}^{b}$ and $h_{k,\phi,n}^{b}$ randomly.
3. Generate the stereophonic mixture $\tilde{x}_2(t)$ as in eqn. (2).
4. Apply the STFT on the mixture signal.
5. E-step: Compute $\hat{p}_{j,fn}$ and $\hat{\boldsymbol{s}}_{fn}$ using (13) and (14).
6. M-step: Compute $\boldsymbol{A}_f$, $\boldsymbol{\Sigma}_{\mathbf{n},f}$, $w_{k,f,\tau}^{y}$, h_{k,ϕ,n_y}^{y}, $w_{k,f,\tau}^{g}$, $h_{k,\phi,n}^{g}$, $\lambda_{k,\phi,n}^{g}$, $w_{k,f,\tau}^{b}$, $h_{k,\phi,n}^{b}$, and $\lambda_{k,\phi,n}^{b}$ using eqns. (22), (24), (34), (35), (28), (29), (30), (31), (32), and (33), respectively.
7. Normalize $w_{k,f,\tau}^{x} = w_{k,f,\tau}^{x} / \sqrt{\sum_{f,k,\tau}\left(w_{k,f,\tau}^{x}\right)^2}$
8. Repeat E- and M-steps, and the normalization until convergence is achieved i.e. rate of cost change is below a prescribed threshold, ψ (e.g. $\psi = -20dB$).
9. Perform inverse STFT with dual synthetic window to estimate $\tilde{g}(t)$, and $\tilde{b}(t)$

Table II

Effects of the identical exemplar on the separation performance

Algorithms	Average SDR (dB)	
	Speech + Music	Speech + Fx
Proposed full-exemplar	4.23	5.96
Proposed semi-exemplar	3.79	5.10

Table III

Effects of the exemplar on the separation performance

Type of Exemplar	Average SDR (dB)
Same Gender	3.06
Different Gender	2.67
Native English Speaker	2.51
Non-Native English Speaker	2.35
Complete sentence	13.75
Missing information sentence	4.95

Table IV

Average SDRs of the 10 mixtures with their different 12 exemplars

Algorithms	Average SDR (dB)	
	Speech + Music	Speech + Fx
Informed excitation-filter channel speech model	-0.74	0.67
Structural GSMM	2.22	2.18
Schmidt's algorithm	1.18	1.34
Proposed semi-exemplar	1.83	2.56
Proposed full-exemplar	2.38	4.04